ESTADÍSTICA ESPAÑOLA
Vol. 37, Núm. 140, 1995, págs. 349 a 362

Nuevo criterio de complejidad utilizando una medida de eficiencia

por
ARANTZA MUNDUATE DEL RIO
Departamento de Física de Materiales
Facultad de Química. Universidad del País Vasco
Apartado 1072
20080 SAN SEBASTIAN
Tfno. 943-216600 (ext.155)

ANTONIO PEREZ PRADOS
Departamento de Métodos Estadísticos
Universidad Pública de Navarra
31006 PAMPLONA

FRANCISCO J. CANO SEVILLA
Departamento de Estadística e Investigación Operativa
Facultad de Matemáticas. Universidad Complutense
28040 MADRID

RESUMEN

Centrado en el campo de los árboles de decisión, este trabajo estudia la forma de selección de un árbol óptimo entre los posibles árboles obtenidos a partir del análisis de un conjunto de datos, utilizando para ello una cantidad criterio que combina linealmente dos medidas de la calidad de un árbol: el error de resustitución y la eficiencia. Analizando el efecto de un proceso de poda en la cantidad criterio se obtiene una sucesión finita de subárboles mínimos óptimamente podados del árbol máximo para los posibles valores del parámetro α de la combinación.

Palabras clave: Arboles de decisión, proceso de poda, eficiencia, error de resustitución.

Clasificación AMS: 62H30

1. INTRODUCCION

El objetivo de las técnicas "cluster" es el agrupamiento de objetos similares. Dentro de estas técnicas se encuentran los árboles de decisión. Un árbol es un conjunto o familia de cluster tal que cualesquiera dos elementos de dicho conjunto o bien son disjuntos o uno incluye el otro. Partiendo de un conjunto de datos para las variables cualitativas $\{V^j\}_{j=1,\ldots,J}$, definidas sobre el conjunto de aprendizaje I, extraído de la población total I, pueden construirse distintos árboles de clasificación. Entre los primeros trabajos relativos al tema se encuentra Hartigan (1975). Debe destacarse en este campo el trabajo de Breiman y otros (1984), al que siguieron otros como Ciampi y otros (1987), Gueguen y Nakache (1988), Ciampi (1991).

Se conocen también diversas medidas que permiten estudiar tanto la calidad de un árbol y su utilidad como predictor de una variable criterio Y que se relaciona con el conjunto de variables $\{V^j\}_{j=1,\ldots,J}$, como los errores de clasificación a que puede dar lugar la utilización del árbol.

Por otra parte, pueden encontrarse en la literatura diversos criterios que permiten determinar un árbol óptimo de acuerdo con una cantidad criterio obtenida a partir de medidas de la calidad del árbol. Así, puede citarse, el criterio de complejidad (Breiman *y otros* , 1984), que utiliza una cantidad criterio que combina linealmente el error de resustitución y la simplicidad del árbol medida a través del número de sus nodos terminales; una generalización de este criterio basada en la contribución de cada nodo interior a la calidad global del árbol se conoce como criterio de contribución (Cuesta, 1989). El criterio del error esperado (Niblett, 1987) utiliza, sin embargo, como medida de calidad, la probabilidad de error al asignar un nuevo ejemplo en una modalidad de Y en un nodo x de T.

En el presente trabajo se propone un nuevo criterio que combina linealmente dos medidas de aspectos distintos de la calidad del árbol. Por una parte el error de resustitución como medida de la capacidad del árbol de clasificar correctamente ejemplos del conjunto I y por otra una medida de la eficiencia del árbol desde el punto de vista de la simplicidad de interpretación del mismo. La medida considerada tiene en cuenta para cada nodo terminal tanto su profundidad como la proporción de elementos del conjunto de aprendizaje I que se sitúan en dicho nodo.

El proceso de selección del árbol es el mismo que el utilizado en el criterio de complejidad (Breiman *y otros*, 1984), pero la medida que hace referencia a la forma

del árbol considera no sólo el número de nodos terminales del mismo sino dos características de estos como son la proporción y profundidad, aumentando con ello la información tenida en cuenta para seleccionar un árbol óptimo en la clase de sub-árboles $\Im=\{T-(T_x)^* : x \text{ nodo interior del árbol } T\}$

2. TERMINOLOGIA Y CONCEPTUALIZACIÓN

2.1. Elementos fundamentales de un árbol

Sea I una población cualquiera de la cual se ha extraído un conjunto de aprendizaje I formado por n elementos, llamados ejemplos, para los cuales se posee información relativa a un conjunto de variables $\{V^j\}_{j=1,...,J}$ a partir de la cual se construye una estructura en árbol T, que es una familia de cluster que presenta distintos niveles de asociación de los ejemplos en estudio, en función del grado de homogeneidad basado en las variables que han sido definidas sobre ellos, verificándose que dos cluster cualesquiera o bien son disjuntos, o uno incluye el otro. Los nodos o vértices del árbol corresponden a subconjuntos de I y se representan genéricamente por x, siendo n_x el número de ejemplos de I situados en dicho nodo x. Cada una de las líneas que partiendo de un vértice llega directamente a otro se llama arco; una sucesión de arcos consecutivos es un camino. Para cada nodo x se dice que x_1 es un nodo sucesor de x si existe un arco que partiendo de x llega a x_1. Por el contrario se dice que x_1 es un nodo generador de x si existe un arco que partiendo de x_1 llega a x. Teniendo en cuenta estos conceptos, entre los nodos de un árbol se distinguen el nodo inicial o nodo raíz x_0 que es aquél que no tiene generador; los nodos terminales que son aquellos que no tienen sucesores. Los nodos del árbol que no son nodos terminales se llaman nodos interiores. El conjunto de nodos interiores del árbol T se representa T^0 y el de nodos terminales $\tilde{T}$. Para un nodo cualquiera x la profundidad es el número de arcos que forman el camino que lo une con el nodo raíz.

2.2. Error de resustitución

Sea Y una variable llamada criterio, que se relaciona con el conjunto de variables $\{V^j\}_{j=1,...,J}$ y cuyo valor se conoce para los elementos del conjunto I. Para el caso de que dicha variable sea cualitativa, que será el analizado en el presente trabajo, sus modalidades serán representadas $\{y_k\}_{k=1,...,C}$. Si T es un árbol construído a partir de la información conocida para los elementos del conjunto de aprendizaje, se designa por T(i) la modalidad de la variable criterio Y que el árbol asigna a un elemento i cualquiera de la población y por Y(i) la modalidad que dicho elemento

presenta. Si ambos valores coinciden, el elemento es clasificado correctamente por el árbol; en caso contrario, se presenta un error de clasificación.

Definición. Se llama error de resustitución de T, $\mathcal{E}R(T)$, a la proporción de ejemplos de I que T clasifica incorrectamente.

Considerando que en cada nodo x se asignan los ejemplos a la modalidad de Y que presenta una mayor proporción, se tiene:

$$e_R(T) = \sum_{x \in T} \mathcal{E}R(x) = \sum_{x \in T} p_x (1 - \max_k p_k^x) \qquad [1]$$

siendo:

p_x la proporción de elementos de I situados en el nodo x.

p_k^x la proporción de la modalidad y_k en el nodo x.

En el caso de que máx $\{p_k^x : k=1,...,c\} = p_r^x = p_s^x$ $r \neq s$ la expresión anterior para $\mathcal{E}R(T)$ es válida tanto si a los elementos del nodo x se les asigna la modalidad y_r como si se les asigna la modalidad y_s; la selección de una entre ambas puede realizarse bien de acuerdo con un criterio que considere el riesgo o pérdida que supone la elección de y_r o y_s al aplicar el árbol para la asignación a un elemento cualquiera de la población total I de una modalidad de Y, o bien eligiendo al azar una de las dos modalidades. La utilización de la función de pérdida permite penalizar los errores de predicción de Y cuando las asignaciones erróneas no son de consecuencias equivalentes en todas las modalidades.

2.3. Proceso de poda

Si x es un nodo cualquiera de T, T_x representa la rama de T generada por x o subárbol engendrado por x en T, es decir el árbol formado por la parte de T que contiene únicamente x y todos sus nodos sucesores hasta llegar a los correspondientes nodos terminales y $(T_x)^*$ representa dicha rama eliminado el nodo x.

Definición. Se llama poda del subárbol T_x de T al hecho de considerar en T el nodo x como terminal eliminando todos los caminos que partiendo de él llegan hasta nodos terminales. El árbol resultante se representa por $T-(T_x)^*$ y se llama subárbol podado de T.

Definición. Si mediante diversas podas se obtiene un conjunto de subárboles podados de T, se llama subárbol podado mínimo, si existe, a aquél que es subárbol podado de todos los demás.

2.4. Método general de selección del árbol óptimo

Se consideran los datos conocidos para los elementos del conjunto de aprendizaje I. A partir de esta información se construye un árbol de forma que en cada nodo terminal todos los elementos que están pertenecen a la misma modalidad de Y. Este árbol se representará $T_{máx}$. Mediante podas sucesivas en el mismo, se obtiene una sucesión de subárboles podados que se comparan de acuerdo con una cantidad criterio elegida previamente. De esta forma podrá seleccionarse el mejor árbol, árbol óptimo, para unas condiciones determinadas.

Según la definición dada para $\mathcal{E}R(T)$ se verifica: $\mathcal{E}R(T_{máx})=0$

Definición. Se dice que T' es un subárbol óptimamente podado de T si el valor de la cantidad criterio a él asociado es el mejor valor entre los correspondientes a todos los subárboles podados de T.

Definición. Se dice que un árbol es subárbol mínimo óptimamente podado de T si:

1. Es óptimamente podado de T y

2. Es subárbol podado de todos los demás que son óptimamente podados.

Si la cantidad criterio combina linealmente el error de resustitución del árbol con una medida de la calidad del mismo, en la forma $C_\alpha(T)=\mathcal{E}R(T)+\alpha M(T)$, el subárbol mínimo óptimamente podado corresponde al mínimo valor de la cantidad criterio y depende del parámetro α de la combinación lineal. Se representa por $T(\alpha)$.

Así, el criterio de complejidad, desarrollado por Breiman (1984), combina el error de resustitución y la simplicidad del árbol medida a través del número de sus nodos terminales, en la forma: $C_\alpha(T)=\mathcal{E}R(T)+\alpha \text{card}T$, comprobándose por inducción, que partiendo de un árbol T cualquiera, para cada valor del parámetro α de la combinación lineal, existe uno y sólo un subárbol mínimo óptimamente podado de T, cuya construcción puede obtenerse partiendo de las ramas primarias del árbol, considerándolas como nuevos árboles y repitiendo el proceso hasta llegar a nodos cuyos nodos sucesores sean nodos terminales. También mediante el método de inducción y teniendo en cuenta la forma de construcción de $T(\alpha)$ se comprueba que dados dos valores α_1 y α_2 del parámetro α, si $\alpha_1 \leq \alpha_2$, el subárbol mínimo óptimamente podado de T correspondiente a α_2 es podado del correspondiente a α_1. Por otra parte, la variación en la cantidad criterio por efecto de un proceso de poda viene dada por:

$$\Delta_x C_\alpha(T)=C_\alpha(T)-C_\alpha(T-(T_x)^*)=\mathcal{E}R(T_x)-\mathcal{E}R(x)+\alpha(\text{card}T_x-1)$$

y consta por lo tanto de dos términos, el primero de ellos $\mathcal{ER}(T_x)$-$\mathcal{ER}(x)$ toma valores negativos o nulos, y el segundo $\alpha(\text{card}T_x-1)$ es positivo salvo en el caso $\alpha=0$.

3. NUEVO CRITERIO UTILIZANDO MEDIDAS DE EFICIENCIA

3.1. Conceptos y resultados fundamentales

Dado un árbol cualquiera T se define la cantidad criterio $E_\alpha(T)$ por:

$$E_\alpha(T) = \mathcal{ER}(T) + \alpha M(T) \qquad [2]$$

donde $\mathcal{ER}(T)$ es el error de resustitución [1], y $M(T)$ es la eficiencia, esto es una medida de la calidad del árbol según la facilidad de su interpretación, definida por:

$$M(T)=\sum_{x\in T} p_x h_x \qquad [3]$$

donde p_x y h_x son respectivamente la proporción y profundidad del nodo x, y α es un valor constante positivo o nulo que penaliza el valor de la eficiencia del árbol en el criterio.

Operando puede escribirse:

$$E_\alpha(T)=\sum_{x\in T} \mathcal{ER}(x) + \alpha \sum_{x\in T} p_x h_x = \sum_{x\in T} (\mathcal{ER}(x) + \alpha p_x h_x)$$

Y teniendo en cuenta [1], se llega a:

$$E_\alpha(T)=\sum_{x\in T} p_x [1 - \max_k p_k^x + \alpha h_x] = 1+\sum_{x\in T} p_x [a h_x - \max_k p_k^x]$$

Si se consideran los dos árboles extremos construídos a partir de los datos para el conjunto de aprendizaje I, esto es el formado por el nodo raíz x_0 únicamente y el árbol $T_{máx}$ obtenido con el criterio de que cada nodo terminal contiene elementos de la misma modalidad de Y, se tiene que:

$$E_\alpha(x_0) = 1-\max_k p_k^{x_0} = \mathcal{ER}(x_0)$$

$$E_\alpha(T_{máx}) = \alpha M(T_{máx})$$

Propiedades

1. Sea x un nodo interior cualquiera de un árbol T y T_x la rama engendrada correspondiente, entonces:

$$E_\alpha(T_x) = \mathcal{E}R(T_x) + \frac{\alpha}{p_x}\sum_{y\in T_x} p_y h_y - \alpha h_x$$

Demostración

La cantidad criterio para la rama T_x es:

$$E_\alpha(T_x) = \mathcal{E}R(T_x) + \alpha M(T_x) = \mathcal{E}R(T_x) + \alpha \sum_{y\in T_x} p'_y h'_y$$

donde p'_y, h'_y representan la proporción y profundidad del nodo y respecto de la rama engendrada T_x. Por tanto:

$$\sum_{y\in T_x} p'_y h'_y = \sum_{y\in T_x} \frac{n_y}{n_x}(h_y - h_x) = \sum_{y\in T_x} \frac{n}{n_x}\frac{n_y}{n} h_y - h_x = \frac{1}{p_x}\sum_{y\in T_x} p_y h_y - h_x$$

Luego:

$$E_\alpha(T_x) = \mathcal{E}R(T_x) + \frac{\alpha}{p_x}\sum_{y\in T_x} p_y h_y - \alpha h_x$$

2. Si se obtiene el árbol T' podando T en el nodo x, la cantidad criterio correspondiente a T' toma el siguiente valor:

$$E_\alpha(T') = E_\alpha(T) - \mathcal{E}R(T_x) - \alpha \sum_{y\in T_x} p_y h_y + \mathcal{E}R(x) + \alpha p_x h_x$$

Demostración

$$E_\alpha(T') = E_\alpha(T-(T_x)^*) = \mathcal{E}R(T-(T_x)^*) + \alpha M(T-(T_x)^*) =$$

$$= \mathcal{E}R(T) - \mathcal{E}R(T_x) + \mathcal{E}R(x) + \alpha\left[M(T) - a_{in}\sum_{y\in T_x} p_y h_y + p_x h_x\right] =$$

$$= E_\alpha(T) - \mathcal{E}R(T_x) - \alpha \sum_{y\in T_x} p_y h_y + \mathcal{E}R(x) + \alpha p_x h_x$$

3. La variación obtenida en la cantidad criterio al pasar del árbol T a su árbol podado T' obtenido al considerar $x \in T^0$ como nodo terminal es:

$$\Delta_x E_\alpha(T) = \mathcal{E}R(T) - \mathcal{E}R(x) + \alpha p_x M(T_x) \quad [4]$$

Demostración

$$\Delta_x E_\alpha(T) = E_\alpha(T) - E_\alpha(T') =$$

$$= E_\alpha(T) - \left[E_\alpha(T) - \mathcal{E}R(T_x) - \alpha \sum_{y \in T_x} p_y h_y + \mathcal{E}R(x) + \alpha p_x h_x\right]$$

Y teniendo en cuenta el resultado obtenido en la propiedad 1

$$\Delta_x E_\alpha(T) = \mathcal{E}R(T_x) - \mathcal{E}R(x) + p_x \left[\frac{1}{p_x}\alpha \sum_{y \in T_x} p_y h_y - \varepsilon h_x\right] =$$

$$= \mathcal{E}R(T_x) - \mathcal{E}R(x) + \alpha p_x M(T_x)$$

Resultado semejante al del criterio de complejidad, es decir, la variación en la cantidad criterio consta de dos términos, el primero de ellos $\mathcal{E}R(T_x) - \mathcal{E}R(x)$ es idéntico al que corresponde al criterio de complejidad, y es negativo; el segundo contiene la parte correspondiente a la medida de calidad y es positivo. El objetivo para la obtención de $T(\alpha)$ será tomar como terminales aquellos nodos que produzcan un subárbol podado cuya cantidad criterio asociada sea mínima y por tanto el valor de $\Delta_x E_\alpha(T)$ sea máximo.

3.2. Sucesión de árboles según los valores de "α"

Notemos por $T'_x = T - (T_x)^*$

Propiedades fundamentales

1. Si x_2 es un sucesor de x_1 se verifica:

i) Si $\mathcal{E}\boldsymbol{R}(T'_{x_1}) = \mathcal{E}\boldsymbol{R}(T'_{x_2}) \rightarrow E_\alpha(T'_{x_1}) \leq E_\alpha(T'_{x_2}) \quad \forall\alpha\geq 0$

ii) Si $\boldsymbol{M}(T'_{x_1}) = \boldsymbol{M}(T'_{x_2}) \rightarrow E_\alpha(T'_{x_1}) \geq E_\alpha(T'_{x_2}) \quad \forall\alpha\geq 0$

iii) Si $\mathcal{E}\boldsymbol{R}(T'_{x_1}) \neq \mathcal{E}\boldsymbol{R}(T'_{x_2})$ y $\boldsymbol{M}(T'_{x_1}) \neq \boldsymbol{M}(T'_{x_2}) \rightarrow \exists\alpha'>0$ tal que

$$E_\alpha(T'_{x_2}) < E_\alpha(T'_{x_1}) \quad \forall\alpha<\alpha'$$

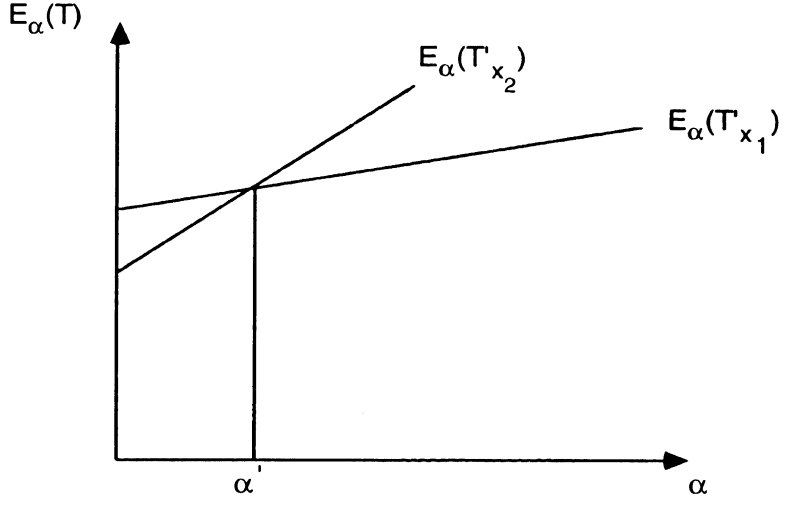
$$E_\alpha(T'_{x_1}) < E_\alpha(T'_{x_2}) \quad \forall\alpha>\alpha'$$

$$E_\alpha(T'_{x_1}) = E_\alpha(T'_{x_2}) \quad \text{si } \alpha=\alpha'$$

Lo cual se deduce del hecho de que $E_\alpha(T)$ es lineal en α,

$$\mathcal{E}\boldsymbol{R}(T'_{x_1}) \geq \mathcal{E}\boldsymbol{R}(T'_{x_2}) \quad \text{y} \quad \boldsymbol{M}(T'_{x_1}) \leq \boldsymbol{M}(T'_{x_2})$$

así iii) corresponde a la gráfica siguiente:



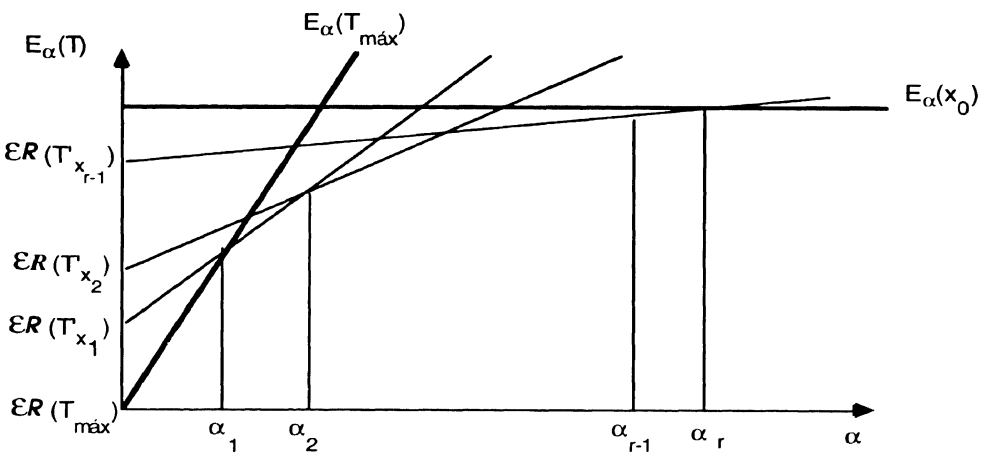
2. La función $F(\alpha)$ = Mínimo $\{ E_\alpha(T') : T' \in \mathfrak{I}\}$, donde $\mathfrak{I}$ es la clase de subárboles podados T'_x y x es un nodo interior de $T_{máx}$ es cóncava y lineal a trozos.

Esta propiedad se deduce del hecho de ser la envolvente inferior de un número finito de funciones lineales y de la propiedad anterior.

Si se ordenan los subárboles T'_x por orden creciente de $\mathcal{ER}(T'_x) = E_\alpha(T'_x)$ para $\alpha=0$, se puede comprobar que:

$$F(\alpha) = \begin{cases} E_\alpha(T_{máx}) & 0 \le \alpha \le \alpha_1 \\ E_\alpha(T'_{x_1}) & \alpha_1 \le \alpha \le \alpha_2 \\ E_\alpha(T'_{x_2}) & \alpha_2 \le \alpha \le \alpha_3 \\ \cdot & \\ \cdot & \\ \cdot & \\ E_\alpha(T'_{x_{r-1}}) & \alpha_{r-1} \le \alpha \le \alpha_r \\ E_\alpha(x_0) & \alpha_r \le \alpha \end{cases}$$

ya que la función $F(\alpha)$ tiene la representación gráfica:



Método de obtención de la secuencia de subárboles mínimos óptimamente podados del árbol máximo.

Para cada valor α el subárbol mínimo óptimamente podado existe y es único.

Si $\alpha=0$ se tiene que $E_\alpha(T)=\mathcal{E}R(T)$. En consecuencia, los subárboles óptimamente podados de $T_{máx}$ serán todos aquellos que corresponden al mínimo valor del error, y por lo tanto que tienen un error de resustitución cero.

Por otra parte, aunque α pueda tomar cualquier valor real y positivo, el número de subárboles podados obtenidos a partir de $T_{máx}$ es evidentemente finito.

Como consecuencia, a medida que α crece se obtiene una secuencia finita de subárboles de $T_{máx}$, T_0 ,T_1 ,T_2 ,...,T_r . Además, por la propia finitud del número de subárboles de la sucesión si $T(\alpha_1)$ es el mínimo subárbol óptimamente podado para el valor α_1, seguirá siéndolo a medida que α crece hasta llegar a otro valor α_2 que dará lugar a un nuevo subárbol $T(\alpha_2)$; este proceso se repite para todo el crecimiento de α. Por idénticas razones a las correspondientes al criterio de complejidad cada uno de los árboles de la sucesión será subárbol podado del anterior, Munduate (1993).

El proceso es, según todo lo anterior, equivalente al correspondiente al criterio de complejidad.

Algoritmo de obtención

Paso 1: Se inicia el proceso con $T_0=T_{máx}$

Paso 2: Si $E_\alpha(T_{i-1})$ es el valor de la cantidad criterio para $T_{i-1}=T(\alpha_{i-1})$, subárbol mínimo óptimamente podado obtenido en el paso i-1, se calcula: α_i = Mín $\{\alpha: \Delta_x E_\alpha(T_{i-1})=0 \quad x\in T^0_{i-1}\}$, entonces : $T_i=T'_{x_i}$ siendo x_i un nodo interior del árbol T_{i-1} en el que se ha alcanzado el mínimo anterior.

Fin: Se repite el paso 2 hasta que $T_r=T(\alpha_r)$ sea el árbol formado sólo por el nodo raíz.

3.3. Selección de un árbol según los valores de "α"

Si el valor del parámetro α que penaliza la medida de la eficiencia del árbol en $E_\alpha(T)$ es conocido, el árbol óptimo de acuerdo con el nuevo criterio de complejidad, obtenido por el método anterior existe y es único. El problema de la selección se plantea cuando α no queda previamente fijado, caso en el cual se conoce una

sucesión de "r" subárboles mínimos óptimamente podados de $T_{máx}$ con sus valores de α asociados.

Pueden existir condiciones que restrinjan los posibles valores de α o impongan condiciones a los árboles óptimos. En general, pueden presentarse las siguientes situaciones:

1. De acuerdo con la aplicación a la que se destina el árbol, los valores del parámetro α de $E_\alpha(T)$ deben verificar algunas condiciones. Según éstas puede ocurrir que parte de los elementos de la sucesión de árboles $T_0, T_1, \ldots T_r$ quede eliminada o incluso que únicamente exista un subárbol mínimo óptimamente podado de $T_{máx}$. En este último caso la selección queda directamente realizada. Si, por el contrario éste no es único se procede a la elección entre los árboles de la subsecuencia obtenida.

2. La profundidad del árbol no debe exceder un valor H concreto. En la práctica esta situación puede considerarse equivalente a la anterior, ya que al crecer los valores de α se van realizando sucesivas podas en los árboles disminuyendo la profundidad de los mismos, por lo que esta restricción puede que haga posibles únicamente valores de α a partir de uno dado, con lo cual la situación es equivalente a la anterior.

3. No existe limitación previa para los posibles valores de α ni para la profundidad del árbol, en consecuencia, de partida cualquiera de los árboles de la sucesión $T_0, T_1, \ldots T_r$ es válido como subárbol mínimo óptimamente podado de $T_{máx}$. La selección del árbol se basa en la elección de aquél que dé lugar a un valor mínimo del error de clasificación estimado. Esta estimación puede realizarse bien a través de un conjunto test o bien mediante validación cruzada, (Goodman *y otros*, 1954; Breiman *y otros* , 1984)

4. CONCLUSIONES

De acuerdo con los resultados anteriores, combinando linealmente el error de resustitución y la eficiencia de un árbol para la formación de la cantidad criterio se propone un método para obtener una sucesión de árboles, subárboles mínimos óptimamente podados del árbol máximo, dependientes del coeficiente α de la combinación lineal.

La variación entre la cantidad criterio correspondiente al árbol T y al árbol $T'=T-(T_x)^*$ depende por una parte de la diferencia entre el error de resustitución del nodo x y de su rama engendrada T_x, y por otra de la eficiencia de esta rama y de la proporción de elementos de l situados en el nodo x. Cualquier proceso de poda da lugar a un aumento tanto del error de resustitución como de la eficiencia, lo cual supone una reducción de la calidad del árbol desde el punto de vista de su capaci-

dad de clasificar correctamente nuevos elementos de la población total I y, sin embargo, un aumento de su eficiencia.

Para la mejor selección de un árbol entre los de la sucesión de subárboles mínimos óptimamente podados será fundamental una conveniente determinación del parámetro α de la cantidad criterio para la aplicación concreta a la que se destina el árbol.

REFERENCIAS

BREIMAN,L., FRIEDMAN,J.H., OLSHEN,R.A., STONE,C.J., (1984). «Classification and Regression Trees». *The Wadsworth Statistics / Probability Series.*

CIAMPI,A., CHANG,C.H., HOGG,S., MCKINEEY,S., (1987). «Recursive partition: A versatile method for exploratory data analysis in biostatistics». *In Proceedings from Joshi Festschrift, G. Umphrey (ed),* pp. 23-50. Amsterdam: North-Holland.

CIAMPI,A., (1991). «Generalized Regression Trees». *Computational Statistics and Data Analysis,* 12, nº1, pp. 57-78.

CUESTA, P. (1989). «Inducción en bancos de datos cualitativos». *Tesis Doctoral. Facultad de Matemáticas. Universidad Complutense de Madrid.*

GOODMAN,L., KRUSKAL,W., (1954). «Measures of Association for Cross Classifications». *JASA*,49; pp. 732-764.

GUEGUEN,A., NAKACHE,J.P., (1988). «Methode de discrimination basée sur la construction d'un arbre de décision binaire». *Revue de statistique appliquée*, 36; pp. 19-38.

Hartigan,j., (1975). «Clustering Algorithms». *Wiley and Sons.* New York:.

MUNDUATE,A., (1993). «Cuestiones notables en la construcción y comparación de árboles de decisión». *Tesis Doctoral. Departamento de Métodos Estadísticos. Universidad Pública de Navarra.*

NIBLETT, T. (1987). «Constructing Decision Trees in Noisy Domains». *Progress in Machine Learning. Sigma Press.*

NEW COMPLEXITY CRITERIUM USING AN EFFICIENCY MEASURE

SUMMARY

In this work, dealing with a criterium for selecting the optimum tree obtained among the possible trees which are deduced from the analysis of a set data. For this purpose, a criterium amount which linearly combines two measures related with the quality of tree is used. Such two measures were the resubstitution error one and the efficiency. Analyzing the effect of a pruning process on the criterium amount, a finite succession of minimum subtrees can be obtained. These subtrees are optimally pruned from the maximum tree according to the parameter of the combination from the above mentioned measures.

Keywords: Decision trees, pruning process, efficiency, resubstitution errror.